

**SYSTEMS AND METHODS FOR TRANSFORMING DATA IN
BUFFER MEMORY WITHOUT UNNECESSARILY COPYING
DATA TO ADDITIONAL MEMORY LOCATIONS**

Cross-Reference

[0001] This application is a continuation-in-part of application Serial Number 10/391,726, filed March 18, 2003 and entitled “SYSTEMS AND METHODS FOR SCHEDULING DATA FLOW EXECUTION BASED ON AN ARBITRARY GRAPH DESCRIBING THE DESIRED DATA FLOW”, the contents of which are herein incorporated by reference.

Technical Field

[0002] The present invention relates generally to database systems and, more particularly, to systems and methods for transforming data in a single buffer location without unnecessarily copying said data to additional memory locations.

Background

[0003] A relational database is a collection of related data that can be represented by two-dimensional tables of columns and rows wherein information can be derived by performing set operations on the tables, such as join, sort, merge, and so on. The data stored in a relational database is typically accessed by way of a user-defined query that is constructed in a query language such as Structured Query Language (SQL).

[0004] Often it is useful to extract data from one or more sources, transform the data into some more useful form, and then load the results to a separate destination. A data warehouse, for example, is a central repository for all or significant parts of the data that an entity's various business systems collect and store (often in separate databases), the purpose of the data warehouse being to support data mining, decision support systems (DSS), and other data actions. Data from various sources is selectively extracted and organized on the data warehouse

database for use by analytical applications and user queries. Data warehousing emphasizes the capture of data from diverse sources for useful analysis and access.

[0005] In the context of a data warehousing, and more generally for managing databases, extract-transform-load (ETL) refers to three separate functions of obtaining, processing, and storing data. The extract function reads data from a specified source database and extracts a desired subset of data. The transform function works with the acquired data—using rules or lookup tables, or creating combinations with other data—to convert it to the desired state as defined by the specific ETL tool. The load function is used to write the resulting data (either all of the subset or just the changes) to a destination database. Various and diverse ETL tools can be used for many purposes, including populating a data warehouse, converting a database of a specific type into a database of another type, or migrating data from one database to another. However, when transforming data, a copy of the data extracted from the source and resident in memory (the primary copy) is again copied to a separate memory location, manipulated, and then copied back to the primary copy location. This unnecessary copying is both inefficient and can lead to data inconsistencies. What is needed in the art is a means for directly transforming a single buffered copy of data in memory without further copying of said data.

Summary

[0006] Various embodiments of the present invention are directed to a unique memory management scheme utilized by the data transformation pipeline (DTP) whereby data extracted from an external source is placed in a memory buffer and is then manipulated by the components without the need for copying the data to any other location in memory. While logically the data moves from component to component in the data flow executions (DFE) for the various embodiments described herein, the data does not in fact change locations. Instead, the data resides in the buffer and is operated upon by a series of components that, in turn, access the data in the buffer via pointers and manipulate same.

Brief Description Of The Drawings

[0007] The foregoing summary, as well as the following detailed description of preferred embodiments, is better understood when read in conjunction with the appended

drawings. For the purpose of illustrating the invention, there is shown in the drawings exemplary constructions of the invention; however, the invention is not limited to the specific methods and instrumentalities disclosed. In the drawings:

[0008] Fig. 1 is a block diagram representing a computer system in which aspects of the present invention may be incorporated;

[0009] Fig. 2A is schematic diagram representing a network in which aspects of the present invention may be incorporated;

[0010] Fig. 2B is a diagram representing tables in an exemplary database;

[0011] Fig. 3 is an architecture of an exemplary database management system;

[0012] Fig. 4 is a network of database systems depicting the logical flow of data;

[0013] Fig. 5 is a diagram showing an ETF-based transformation of data as it moves between databases;

[0014] Fig. 6A is an illustration of the functional structure of one embodiment of a data transformation pipeline;

[0015] Fig. 6B is an illustration of data flow execution for the embodiment illustrated in of Fig. 6A.

[0016] Fig. 7A is a diagram of a graph that a user might develop using the graphical user interface of one embodiment of the data transformation pipeline.

[0017] Fig. 7B is the optimized data flow execution for the user-defined graph illustrated in Fig. 7A.

[0018] Fig. 7C illustrates the data stored in the buffer of the data transformation pipeline immediately after extraction and storage.

[0019] Fig. 7D illustrates the data stored in the buffer of the data transformation pipeline immediately after the data is transformed (sorted).

[0020] Fig. 7E illustrates the data stored in the buffer of the data transformation pipeline immediately after the data is split into two different data groups.

[0021] Fig. 8A illustrates an exemplary data structure for a phone book pertaining to one embodiment of the present invention.

[0022] Fig. 8B illustrates the cumulative offsets, from the beginning of a row, for starting points of each data column within each row of the data structure of Fig. 8A.

[0023] Fig. 9 illustrates an exemplary buffer data structure for one embodiment of the present invention.

[0024] Fig. 10A illustrates a split transform of data in a buffer structure pertaining to various embodiments of the present invention.

[0025] Fig. 10B illustrates a hypothetical group of celebrities from a single source that are to be split into two groups, “Winners” and “Losers”, by the split transform of Fig. 10A.

[0026] Fig. 10C illustrates an exemplary buffer data structure for one embodiment of the present invention for splitting the celebrities of Fig. 10B into two groups according to the split transform of Fig. 10A.

Detailed Description

Overview

[0027] The following discussion is directed to a system and method for scheduling data flow execution based on an arbitrary graph describing the desired flow of data from at least one source to at least one destination. The subject matter is described with specificity to meet statutory requirements. However, the description itself is not intended to limit the scope of this patent. Rather, the inventors have contemplated that the claimed subject matter might also be embodied in other ways, to include different elements or combinations of elements similar to the ones described in this document, in conjunction with other present or future technologies. Moreover, where the embodiments described herein describe the invention in connection with row-level access and processing, it should be noted that the invention is by no means limited to row-level access and processing. For example, the invention could be applied on a column basis or a table basis as well.

Computer Environment

[0028] Fig. 1 and the following discussion are intended to provide a brief general description of a suitable computing environment in which the invention may be implemented. Although not required, the invention will be described in the general context of computer executable instructions, such as program modules, being executed by a computer, such as a workstation or server. Generally, program modules include routines, programs, objects, components, data structures and the like that perform particular tasks or implement particular

abstract data types. Moreover, those skilled in the art will appreciate that the invention may be practiced with other computer system configurations, including hand held devices, multi processor systems, microprocessor based or programmable consumer electronics, network PCS, minicomputers, mainframe computers and the like. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

[0029] With reference to Fig. 1, an exemplary system for implementing the invention includes a general purpose computing device in the form of a conventional personal computer 20 or the like, including a processing unit 21, a system memory 22, and a system bus 23 that couples various system components including the system memory to the processing unit 21. The system bus 23 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. The system memory includes read only memory (ROM) 24 and random access memory (RAM) 25. A basic input/output system 26 (BIOS), containing the basic routines that help to transfer information between elements within the personal computer 20, such as during start up, is stored in ROM 24. The personal computer 20 may further include a hard disk drive 27 for reading from and writing to a hard disk, not shown, a magnetic disk drive 28 for reading from or writing to a removable magnetic disk 29, and an optical disk drive 30 for reading from or writing to a removable optical disk 31 such as a CD ROM or other optical media. The hard disk drive 27, magnetic disk drive 28, and optical disk drive 30 are connected to the system bus 23 by a hard disk drive interface 32, a magnetic disk drive interface 33, and an optical drive interface 34, respectively. The drives and their associated computer readable media provide non-volatile storage of computer readable instructions, data structures, program modules and other data for the personal computer 20. Although the exemplary environment described herein employs a hard disk, a removable magnetic disk 29 and a removable optical disk 31, it should be appreciated by those skilled in the art that other types of computer readable media which can store data that is accessible by a computer—such as magnetic cassettes, flash memory cards, digital video disks, Bernoulli cartridges, random access memories (RAMs), read only memories (ROMs) and the like—may also be used in the exemplary operating environment. Further, as

used herein, the term computer readable medium includes one or more instances of a media type (e.g., one or more floppy disks, one or more CD-ROMs, etc.).

[0030] A number of program modules may be stored on the hard disk, magnetic disk 29, optical disk 31, ROM 24 or RAM 25, including an operating system 35, one or more application programs 36, other program modules 37 and program data 38. A user may enter commands and information into the personal computer 20 through input devices such as a keyboard 40 and pointing device 42. Other input devices (not shown) may include a microphone, joystick, game pad, satellite disk, scanner or the like. These and other input devices are often connected to the processing unit 21 through a serial port interface 46 that is coupled to the system bus, but may be connected by other interfaces, such as a parallel port, game port or universal serial bus (USB). A monitor 47 or other type of display device is also connected to the system bus 23 via an interface, such as a video adapter 48. In addition to the monitor 47, personal computers typically include other peripheral output devices (not shown), such as speakers and printers.

[0031] The personal computer 20 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 49. The remote computer 49 may be another personal computer, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the personal computer 20, although only a memory storage device 50 has been illustrated in Fig. 1. The logical connections depicted in Fig. 1 include a local area network (LAN) 51 and a wide area network (WAN) 52. Such networking environments are commonplace in offices, enterprise wide computer networks, Intranets and the Internet.

[0032] When used in a LAN networking environment, the personal computer 20 is connected to the local network 51 through a network interface or adapter 53. When used in a WAN networking environment, the personal computer 20 typically includes a modem 54 or other means for establishing communications over the wide area network 52, such as the Internet. The modem 54, which may be internal or external, is connected to the system bus 23 via the serial port interface 46. In a networked environment, program modules depicted relative to the personal computer 20, or portions thereof, may be stored in the remote memory storage device. It will be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be used.

Network Environment

[0033] Fig. 2A illustrates an exemplary network environment in which the present invention may be employed. Of course, actual network and database environments can be arranged in a variety of configurations; however, the exemplary environment shown here provides a framework for understanding the type of environment in which the present invention operates.

[0034] The network may include client computers 20a, a server computer 20b, data source computers 20c, and databases 70, 72a, and 72b. The client computers 20a and the data source computers 20c are in electronic communication with the server computer 20b via communications network 80, e.g., an Intranet. Client computers 20a and data source computers 20c are connected to the communications network by way of communications interfaces 82. Communications interfaces 82 can be any one of the well-known communications interfaces such as Ethernet connections, modem connections, and so on.

[0035] Server computer 20b provides management of database 70 by way of database server system software, described more fully below. As such, server 20b acts as a storehouse of data from a variety of data sources and provides that data to a variety of data consumers.

[0036] In the example of Fig. 2A, data sources are provided by data source computers 20c. Data source computers 20c communicate data to server computer 20b via communications network 80, which may be a LAN, WAN, Intranet, Internet, or the like. Data source computers 20c store data locally in databases 72a, 72b, which may be relational database servers, excel spreadsheets, files, or the like. For example, database 72a shows data stored in tables 150, 152, and 154. The data provided by data sources 20c is combined and stored in a large database such as a data warehouse maintained by server 20b.

[0037] Client computers 20a that desire to use the data stored by server computer 20b can access the database 70 via communications network 80. Client computers 20a request the data by way of SQL queries (e.g., update, insert, and delete) on the data stored in database 70.

Database Architecture

[0038] A database is a collection of related data. In one type of database, a relational database, data is organized in a two-dimensional column and row form called a table. Fig. 2B illustrates tables such as tables 150, 152, and 154 that are stored in database 72a. A relational

database typically includes multiple tables. A table may contain zero or more records and at least one field within each record. A record is a row in the table that is identified by a unique numeric called a record identifier. A field is a subdivision of a record to the extent that a column of data in the table represents the same field for each record in the table.

[0039] A database typically will also include associative structures. An example of an associative structure is an index, typically, but not necessarily, in a form of B-tree or hash index. An index provides for seeking to a specific row in a table with a near constant access time regardless of the size of the table. Associative structures are transparent to users of a database but are important to efficient operation and control of the database management system. A database management system (DBMS), and in particular a relational database management system (RDBMS) is a control system that supports database features including, but not limited to, storing data on a memory medium, retrieving data from the memory medium and updating data on the memory medium.

[0040] As shown in Fig. 2B, the exemplary database 72a comprises employee table 150, department table 152, and sysindexes table 154. Each table comprises columns 156 and rows 158 with fields 160 formed at the intersections. Exemplary employee table 150 comprises multiple columns 158 including empl_id, empl_name, empl_salary, and dept_id. Columns 158 in department table 152 include dept_id, dept_name, and dept_location. Sysindexes table 154 contains information regarding each table in the database.

[0041] Generally, data stored in a relational database is accessed by way of a user-defined query that is constructed in a query language such as SQL. Typically, for any given SQL query there are numerous procedural operations that need be performed on the data in order to carry out the objectives of the SQL query. For example, there may be numerous joins and table scans that need to be performed so as to accomplish the desired objective.

[0042] As noted, control and management of the tables is maintained by a DBMS, e.g., a RDBMS. An exemplary SQL Server RDBMS architecture 90 is graphically depicted in Fig. 3. The architecture comprises essentially three layers. Layer one provides for three classes of integration with the SQL Server, comprising: (1) a SQL Server Enterprise Manager 92 that provides a common environment for managing several types of server software in a network and provides a primary interface for users who are administering copies of SQL Server on the network; (2) an Applications Interface 93 that allows integration of a server interface into user

applications such as Distributed Component Object Modules (DCOM); and (3) a Tools Interface 94 that provides an interface for integration of administration and configuration tools developed by Independent Software Vendors (ISV).

[0043] Layer two opens the functionality of the SQL server to other applications by providing three application programming interfaces (API): SQL Namespace 95, SQL Distributed Management Objects 99, and Data Transformation Services 100. A user interface 91 is provided by Wizards, HTML, and so on. SQL Namespace API 95 exposes the user interface (UI) elements of SQL Server Enterprise Manager 92. This allows applications to include SQL Server Enterprise Manager UI elements such as dialog boxes and wizards.

[0044] SQL Distributed Management Objects API 99 abstracts the use of DDL, system stored procedures, registry information, and operating system resources, providing an API to all administration and configuration tasks for the SQL Server.

[0045] Distributed Transformation Services API 100 exposes the services provided by SQL Server to aid in building data warehouses and data marts. As described more fully below, these services provide the ability to transfer and transform data between heterogeneous OLE DB and ODBC data sources. Data from objects or the result sets of queries can be transferred at regularly scheduled times or intervals, or on an ad hoc basis.

[0046] Layer three provides the heart of the SQL server. This layer comprises an SQL Server Engine 97 and a SQL Server Agent 96 that monitors and controls SQL Server Engine 97 based on Events 98 that inform SQL Server Agent of the status of the SQL Server Engine 97.

[0047] The Server Engine processes SQL statements, forms and optimizes query execution plans, and so on.

Logical Database Application

[0048] The above description focused on physical attributes of an exemplary database environment in which the present invention operates. Fig. 4 logically illustrates the manner in which data moves among a number of database servers, which may simultaneously be data sources for other database servers, to the destination database. Here, database server 20b provides management of database 70. Data for database 70 is provided by data sources 72a and 72b, which are managed by database servers 20c' and 20c, respectively. Significantly, database 20c' gathers data from databases 72c and 72d, which are managed by servers 20d. Thus,

database 70 is fed directly with data from databases 72a and 72b and indirectly with data from databases 72c and 72d.

[0049] In the exemplary system of this figure, data from database 72c moves through database 72a and then on to database 70. Along the way, the data may also undergo transformation. This example illustrates the general concept how data movement may comprise several hops in order for such data to actually reach the database server of interest. Those skilled in the art will recognize that many other combinations of movement and transformation of data is possible.

[0050] Fig. 5 illustrates a transformation using multiple ETL tools. In this exemplary transfer, data is merged from two different tables that reside in two different databases into a third table residing in a third database. For example, table 150 resides in database 72a whereas table 149 resides in database 72b. The tables are merged into a third table 151 that is maintained in database 70.

[0051] Although both tables 149, 150 contain similar information, it is not in an identical format. As a result, the data must be transformed by separate ETL tools into the format of table 151. For example, table 150 maintains a column empl_name that contains employee names as first name followed by last name; whereas, table 149 maintains a column name that contains employee names as last name followed by first name. Table 151 contains employee names in the form of table 150. In order for the name columns of table 149 to be inserted into the empl_name column of table 151, the name must be converted to the proper form. Similarly, table 149 does not contain dept_id information.

[0052] The above example illustrates that data moving between databases may need to be transformed in some manner before insertion into the target database. However, using separate ETL tools to achieve each transformation is inefficient. In Fig. 5, for example, transformation application 204 (one ETL tool) transforms the data of table 149 into proper form for table 151 and transformation application 202 (a separate ETL tool) transforms the data of table 150 into proper form for table 151.

Data Transfer Service and Data Transfer Pipeline

[0053] A data transformation system (DTS)—various embodiments of which are described in, and are the central subject matter of, parent application Serial Number 10/391,726,

filed March 18, 2003 and entitled “SYSTEMS AND METHODS FOR SCHEDULING DATA FLOW EXECUTION BASED ON AN ARBITRARY GRAPH DESCRIBING THE DESIRED DATA FLOW”—comprises a capability to receive data from a data source (such as a data retrieval system that receives data from a source), a data destination and a capability to store transformed and or non-transformed data therein (a destination data storage system to store data), and a data transformation pipeline (DTP) that constructs complex end-to-end data transformation functionality (data flow executions or DFEs) by pipelining data flowing from one or more sources to one or more destinations through various interconnected nodes (that, when instantiated, become components in the pipeline) for transforming the data as it flows by (where the term transforming is used herein to broadly describe the universe of interactions that can be conducted to, with, by, or on data). Each component in the pipeline possesses specific predefined data transformation functionality, and the logical connections between components define the data flow pathway in an operational sense.

[0054] One solution to the efficiency problem of traditional ETL-based transformations is the use of the data transformation pipeline (DTP), the functional structure of an exemplary one of which is illustrated in Fig. 6A. In this particular embodiment, the DTP 302 comprises a graphical user interface (GUI) 304 that enables a user 300 (represented here as a PC client computer) to develop a complex end-to-end data transformation function (a data flow execution or DFE) by graphically describing and representing a desired data flow from one or more sources to one or more destinations through various interconnected nodes (a graph). Each node in the graph represents specific predefined data transformation functionality that is offered by uninstantiated component objects 370 residing in a component library 316, and connections between the nodes as drawn by the user 300 represent the data flow pathway between the components for the graph.

[0055] After the user 300 inputs graph data 306 via the GUI 304, the DTP 302 utilizes a translator 308 to traverse the graph data 306 and to translate the graph into an DFE plan (not shown). Moreover, in this example, the translator 308 works in conjunction with an optimizer subsystem 310 to optimize the simple graph developed by the user 300 into a maximally efficient execution structure by eliminating redundancies, simplifying and enhancing the DFE plan and possibly performing a plethora of other optimizations that are known and appreciated by those of skill in the art. Based on the DFE plan, the scheduler 314 uses its pipeline engine 318 to build

the actual DFE 320 by instantiating appropriate components objects 370 from the component library 316 (as detailed in Fig. 6B). The translator 308 also produces work lists 312 for the scheduler 314 where each work list 312 contains specific work items for the scheduler 314 to control the operation of the DFE 320; moreover, it is important to note that the actual interconnectivity between the various component objects is in effect reflect in the work lists as parameters associated with each work item in each work list. Of particular note, the DTP 302 also comprises a buffer 380 which is utilized by the DFE 320 (described more fully later herein).

[0056] Notwithstanding the name, the scheduler 314 does not schedule work items according to time, but instead the scheduler 314 manages the work lists 312 and the execution of the work items in the lists by the DFE 320. Each work item in a work list is one of five operations that the scheduler 314 uses to control the operation of the DFE 320, the five operations comprising:

- extracting data from a data source
- providing data to a component (for transformation)
- splitting data from a single path onto two or more paths
- merging data from two or more paths into a single path
- providing data to a data destination

[0057] Referring to both Figs. 6A and 6B, the latter of which is a detailed view of the DFE 320, the first operation, extracting data from a data source, is a work item that causes the scheduler 314 to thread/task/program/schedule/etc. (hereinafter, simply to thread) a specific extraction component—for example, extraction component 322—to extract certain data from a certain data source—for example, data source 350—and for the extraction component 322 to logically hold that data to be passed to another component although, in reality, the data is actually stored in a buffer 380. The second operation, providing data to a component, causes the scheduler 314 to thread a specific component—for example, transformation component 326—to transform the data according to the input/output functionality of the component 326. (As described more fully below, in operation the scheduler 314 actually passes a first set of pointers (primary pointers) for the buffer data to the component so that the component can directly access the data in the buffer 380 and transform it without having to copy it.) The third operation, enabling the split of data along two or more paths, is a work item that causes the scheduler 314 to

thread a specialized component—for example, split component 332—to analyze each row of data and, based on a specified criteria, group the data into one of two groups, alpha (first path) or omega (second path), each of which will thereafter logically travel along separate paths of the pipeline during continuing execution of the DFE 320. Moreover, from this point forward, the scheduler 314 treats alpha and omega as distinct and separate data groups. The fourth operation, enabling the merger of data from two or more paths into a single path, is the logical converse of a split that causes the scheduler 314 to thread another specialized component—for example, merge component 330—to merge two distinct and separate data groups into a single data group that travels along a common path in the pipeline during continuing execution of the DFE 320, and from this point forward the scheduler 314 treats the merged data as a single group. The fifth operation, loading data to a data destination, is a work item that causes the scheduler 314 to thread a specific loading component—for example, loading component 332—to load certain data onto a certain data destination—for example, data destination 360. These five operations comprise the general functionality of the scheduler 314, although it is the specific input/output functionality of the uninstantiated component objects 370 that are available to the DTP 302 (and which are threaded to by the five operational elements) that enable the development of complex data transforms.

[0058] As previously alluded to herein above, the DTP 302 has a multitude of uninstantiated component objects categorized in a component library 316, each of which has defined inputs and outputs from which the user can graphically construct complex data transformations (via the pipeline engine 318) by combining the functionality of the components into an DFE 320 in order to achieve a desired end results. The transformation components are similar to a plurality of ETL tools but individually lack the individual functionality of ETL tools to extract and load data (as these tasks are handled by the scheduler in the DTP system through special, non-transformation components such as the extract components 322 and the load components 334). Moreover, all of the components provide black box transformation functionality—that is, components can be developed on a variety of platforms (Java, ActiveX, etc.) because the development platform is irrelevant to the DTP 302 as it (and the user 300) are only concerned about the inputs and outputs for the component—that is, the functional information necessary to use the component as a black box object.

[0059] Referring again to Figs. 6A and 6B, after the DFE 320 is formed by the pipeline engine 318 as described earlier herein, the scheduler begins executing the individual work items in one of the work lists 312, the individual work items of which are textually depicted Fig. 6B. For example, in executing a work list 312, the scheduler might individually thread the extraction components 322 and 324 to extract data from three data sources 350, 352, and 354 to create two data groups (which are stored as two distinct data groups, A and B respectively, in the buffer 380). Upon each completed extraction, the scheduler then threads the appropriate transformation component 326 or 328 to begin transforming the data corresponding to each path (A and B respectively). When component 326 is complete, and presuming that component 326 is complete before component 328, the scheduler recognizes that the next step for data group A (hereinafter A) is to merge with data from a split process 322 and, since that data is not yet available, the scheduler may not yet initiate the thread for the merger component 330. Meanwhile, component 328 completes its transformation of data group B (hereinafter B) and the scheduler 314 then threads the split component 332 to split B according to input parameters specified by the work item. Consequently, B is logically split into B1 and B2, each data group being the output of component 332 along separate paths in the DFE 320. Once the split is complete the scheduler 314 then threads the merger component 330 to merge A and B1. Also, recognizing that the remaining execution of B2 is independent from the continuing execution of A and B1, the scheduler 314 also threads component 324 to transform B2.

[0060] Without further regard to each remaining pathway, and to summarize the rest of the dataflow in the DFE 320 (without explicitly referring to the scheduler 314, the operation of which can be easily implied), A and B1 are merged by component 320 to form AB, which is then transformed by component 326 and thereafter loaded to an external data destination 360 by loading component 332. Meanwhile, B2, having been transformed by component 324, is then transformed by components 328 and 330 in order, and thereafter B2 is loaded to two external data destinations 362 and 364 by loading component 334.

[0061] This exemplary scheduler, including the important translator/optimizer functionality that has been separate in the figures for clarity but which may in fact be scheduler subsystems, performs a critically important role in the DTP. Not only does the scheduler enable a user to describe complex data transformations in a simple graph easily drawn via the GUI interface, and not only does the scheduler (via the translator) map the graph to an DFE plan and

task lists, but it also controls the actual execution of the data flows throughout the actual DFE to ensure consistent functionality for situations such as: pipelining data through a DFE comprising both synchronous and asynchronous components (where the latter requires all data to be inputted before any data can be outputted); keeping data in sequence when necessary while employing parallel processing techniques; load balancing; enabling parallel processing despite the data residing in a single location in memory (as discussed below); and so forth. Consequently, an important consideration for such an exemplary system is the care given to ensuring that the relationship between elements in a graph and the DTPs capabilities are clearly defined and entirely consistent.

Memory Management Scheme

[0062] Various embodiments of the present invention are directed to a unique memory management scheme utilized by the DTP whereby data extracted from an external source is placed in a memory buffer and is then manipulated by the components without the need for copying the data to any other location in memory. While logically the data moves from component to component in the DFE for the various embodiments described herein, the data does not in fact change locations but, instead, the data resides in the buffer and is operated upon by a series of components that, in turn, access the data in the buffer via pointers and manipulate same.

[0063] Consider Figs. 7A, 7B, 7C, 7D, and 7E which collectively illustrate how data is (a) extracted by an extraction component and stored in buffer memory, (b) transformed by a component, and then (c) loaded to a destination from the buffer memory by a loading component. Fig. 7A (with references to other prior figures) illustrates a sample graph 400 for a data transformation specified by a user 300 via the GUI 304. The user 300 for this example has described that data regarding his subordinates employees should be extracted from the corporate database 402, divided into two groups based on sex (male or female) 404, each group then sorted by name 406 and 408 and, finally, each group loaded to the two separate databases 410 and 412.

[0064] Fig. 7B illustrates how the translator 308 and optimizer subsystem 310 (which, again, may in fact be subsystems of the scheduler 314) might translate the graph upon traversal into an DFE plan that pipeline engine 318 of the scheduler 314 would build from the component library. (The connectivity between objects is also reflected in the task list 312 that the translator

would produce.) In this figure, the resultant DFE 420 comprises an extraction component 424 to extract the necessary data from an external data source 422. The data is then, in a logical sense, passed to transformation component 426 that sorts the data by last name. The data is then passed to a split component 428 to divide the data into a male group of data and a female group of data (Male and Female respectively). Male and Female are then loaded to separate external data destinations 434 and 436 by loading components 430 and 432 respectively. As an aside, it is important to note how the translator 308 and optimizer 310 implemented the graph to require only five components to achieve greater optimization (whereas a direct mapping of the graph would have required six components).

[0065] Figs. 7C and 7D illustrate how the data is physically manipulated in the buffer 380 during initial execution of the DFE 320 both before and after the a sort occurs. Here the data is a table 440 comprising ten rows 442 of five columns each 444, each row containing a different record and each column containing a different record field. The data table is stored in the buffer 380 and primary pointers 446 are created and used to point to each row of data in the table 400. (Additional pointer sets 448 and 450 are also created but can be ignored for now as they will be discussed in more detail later herein.) When the DFE passes the data to the first post-extraction component, the sorting component 426, the scheduler actually passes the primary pointers 446 to the transformation component 426, and it is the primary pointers 446 that are sorted by the transformation component 426 by changing which rows each pointer (which are ordered) actually points to, a technique well-known and appreciated by skilled artisans in the relevant art. Again, it is important to note that the data itself has not moved in buffer, nor has it been duplicated or copied to other memory or, in this case, otherwise manipulated.

[0066] Fig. 7E illustrates how the data is physically manipulated in the buffer 380 during continued execution of the DFE 320. Having already been sorted, the data is then passed by the transformation component 426 to the split component 428 by passing the sorted pointers 446 to the split component 428. The split component then split the data into two groups, Male and Female, by pointing to the male rows with secondary pointers 448 and by pointing to the female rows with tertiary pointers 450. (During translation, the translator 308 recognized that a split would occur and, in anticipation of this split, the buffer created two additional sets of pointers 448 and 450 specifically for the functionality described herein.) Hereinafter, Male and Female are treated by the DFE as separate and distinct data groups and the primary pointers 446

are no longer relevant (in which case the scheduler 314, realizing the primary pointers 446 are no longer needed, could in fact destroy these pointers and release their resources back to the system; however, for convenience, the diagram merely shows the primary pointers no longer pointing to the data). Hereafter, the secondary pointers are then passed to loading components 430 and 432 respectively for those components to load their respective data Male and Female to their respective external data destinations 434 and 436. Note that throughout the entire DFE 320, the actual data in the buffer did not have to move, nor would any movement be necessary for a transformation component to modify any data cell (row by column) as it would simply do so by modify the actual data cell in the buffer memory directly.

[0067] Various embodiments of the present invention are directed to a memory management scheme—that is, a buffer system—comprising a plurality of buffer objects, services by which ETL objects can manipulate the data in the buffers, and a buffer manager for managing the buffer objects and services. The buffer manager further comprises a central repository of type information which describes the format of data stored in each possible buffer type. Once a buffer type is registered, the buffer manager can be asked to create an instance of that type of buffer for use by the pipeline and its transforms.

[0068] In certain embodiments, the buffer manager may be written with the memory usage patterns of the pipeline in mind such that, when buffers of a certain type are used within a subset of the pipeline's transformation graph to repeatedly access and release memory, the buffer manager, in order to make memory available rapidly, sets aside unused buffers instead of destroying them since it is likely the memory and its buffer structure will be used again. This approach can avoid problems with fragmentation in the operating system's own heaps.

[0069] Moreover, instead of simply providing ETL objects the ability to manipulate data by calling individual rows—which requires significant overhead in making call after call for each individual row, and where the cost of setting up and cleaning up after each call is very large—various embodiments of the present invention instead provide ETL objects with a buffer of rows for processing, and arranges each row in such a way that all columns within the row are in close proximity within buffer memory to maximize the efficient utilization of the processor's cache.

[0070] In addition, while the memory manager does allow ETL objects to access data using a simple call-by-cell mechanism, this buffer approach also makes it possible for an ETL

object to access buffer data an entire row at a time because the data manager loads the entire row even for a single cell call. In other words, efficiently accessing one row of data with a single function call—regardless of how many columns are in that row, or how wide the total data in the row is—gives a gain in efficiency over making repeated calls for different columns in a single row.

[0071] Since the buffer object itself is offered to each transform (via the passing of pointers), the buffer does not need to be moved throughout memory to be accessed by each object operating in its data—that is, whenever data is requested by an ETL object, a reference (pointer) to the data is provided. The calling ETL object manipulates data through the reference and, when completed, leaves the data residing in the buffer. This approach avoids copying data to another memory location, revising the data, and then copying the data back to the buffer. Likewise, no copying is necessary when loading data into the buffer from a data source, nor is copying required when reading data from the buffer for writing to a data destination, since the memory manager maintains full rows of data for each call, and the data is bound directly into the buffer without first storing it elsewhere in memory for such operations.

Buffer Types and Initialization

[0072] A buffer type is defined by declaring a set of columns that are characteristic for each row of data. Each such column, common to each row, is assigned a data type, a maximum data width (which is important for data types that do not have a fixed or predefined width), and supplemental info corresponding to the specific data type—for example, the maximum stored precision for a decimal number type. Using this typing schema, a buffer type inherently possesses a known (or at least determinable) row “width” which is the sum of the widths of all the individual columns in the row. For example, consider a phone book which has the data structure shown in Fig. 8A; in this example, the maximum width of each row is 170 characters which, in turn, is the sum of the widths of the three columns that comprise each row ($80 + 10 + 80 = 170$).

[0073] To create a buffer of this type, the buffer manager allocates the highest number of complete rows that fit into the buffer’s default memory size. In one embodiment, the default memory size is nominally ten megabytes (10,485,760 bytes), although other sizes are available for other embodiments, and in certain embodiments this size is adjustable either automatically or

manually. Referring again to the structure of Fig. 8A, the buffer manager would allocate a buffer with 61,680 rows—totaling 10,485,600 bytes—for the present data structure. However, if the row size is so great that it's larger than the default memory size, in various embodiments the buffer type will be allocated enough memory for one full row. While moving a single row of data through the pipeline in a buffer is logically expensive, this solution is still preferable to not being able to handle an arbitrary row width of a size greater than the default memory size. An alternative to this single-row method is to consume more memory such that the memory manager allocates sufficient memory from among the total memory available to provide a buffer of multiple rows.

[0074] In any event, the memory allocated for the buffer may then be allocated by the operating system in a single large one-dimensional arrangement. For use in a database-oriented application, this memory is then logically divided into an irregular two dimensional array where the address for any row can initially be determined using simple math—that is, where the start of a row is given by the row's index multiplied by the row size to determine the correct offset for that data row. Then, for a specific data cell within said row, an additional offset, illustrated in Fig. 8B, can then be added to this sum. Therefore, to find the location of a particular cell at row n , column m , the buffer manager can determine the address using the following algorithm:

$$\text{CellAddress}(m, n) = \text{ColumnOffset}[m] + \text{RowStart}[n]$$

[0075] However, in regard to this basic buffer layout in memory, and for certain alternative embodiments of the present invention, the buffer manager may have two freedoms in planning the layout of rows in the buffer; first, the offset of an individual column might not parallel its relative position in the declaration of the type; and, second, dead space might exist between the end of one cell in a row and the beginning of the next. These two freedoms will enable the buffer manager to tune the exact placement of the cells in memory for any of a variety of reasons known and appreciated by those of skill in the art, such as, for example, making maximum utility of the processor's cache.

Reorganizing Rows in a Buffer

[0076] While an initial reference to a row can be found using simple multiplication (as described above), it is often advantageous to logically reorder rows within a buffer. For example, if a buffer is sorted by customer name, the order of rows must necessarily change to reflect the order provided by the sort. However, it is undesirable to copy data for each involved row to a new memory location to reflect the sort results.

[0077] Therefore, for various embodiments of the present invention, the buffer maintains a pointer to each row's initial starting location (the row start array, "primary pointers", or first set of pointers), and these pointers are initialized mathematically for each row n using the following algorithm:

$$\text{RowStart}(n) = \text{BufferStart} + \text{RowSize} * n$$

[0078] Thereafter, when an ETL object requests the buffer manager to swap any rows (such as in the context of reordering the data in a sort), the buffer manager simply exchanges the pointers corresponding to the starts of subject rows in the row start array. For various embodiments of the present invention, the array of pointers to each row's start (the primary pointers) are maintained in the buffer's instance data, the array using one entry for each possible row in the buffer.

[0079] Of course, while the buffer allocates enough memory for several thousand rows upon initialization, the buffer does not actually not use any of the rows immediately but, instead, the memory is allocated to each row as follows: As rows are added to an empty buffer, they fill space at row indexes starting with zero and working their way upward towards the number of rows in the buffer. This total is the active count of rows in the buffer and is maintained in the buffer's instance data. The current number of rows in use is very different than the maximum number of rows in use, which is also stored in the buffer object instance data.

[0080] Fig. 9 illustrates a buffer having a maximum of ten rows, each row having a width of 170 characters (and thus the buffer occupies a total of 1700 bytes of memory). In this example, only seven rows are in use while the remaining three are unused and therefore uninitialized. An array of pointers—the primary pointers—point to the start of each row, and the row count and max rows member data reflect the present state of this buffer. The buffer's type

information is shown in the lower half of the diagram, and the dashed-line arrow from the offset column are intended to help represent the sizing and spacing of columns for the offset of specific columns in each row.

[0081] Many embodiments of the present invention used in pipeline applications will need to split data which starts at a single source but is then divided for separate processing based on some attribute of the data. The buffer manager plays a critical role in efficiently supporting these scenarios by providing additional “virtual buffers” such that, when the pipeline determines how each is to be split (that is, which group it is to belong to), the buffer manager allocates additional space within the buffer for additional pointer arrays (“subsets”) to support the split without copying any of the rows. For example, if a split transform, such as the one shown in Fig. 10A, picks “winners” and “losers” from among rows pertaining to popular culture celebrities, the desired result would be to split the source input into the corresponding two outputs illustrated by Fig. 10B for a hypothetical group of such celebrities. However, instead of copying each row to a separate buffer destination corresponding with each of the two possible outputs, the buffer manager instead uses the aforementioned subsets built into the buffer and each subset is implemented in the buffer by using a new active row count and a new row starts array as shown in Fig. 10C. Although not shown, both the Winners Subset and the Losers Subset arrays are big enough for ten full rows (since, in the extreme case, all ten rows could all be winners or all be losers); however, for convenience, the unused portion of these arrays are not shown. Moreover, while the subsets in this example are mutually exclusive (that is, each row is either a Winner or a Loser but not both nor neither), other embodiments of split functionality may allow a row to belong to both resultant arrays and/or neither resultant array.

Conclusion

[0082] The various techniques described herein may be implemented with hardware or software or, where appropriate, with a combination of both. Thus, the methods and apparatus of the present invention, or certain aspects or portions thereof, may take the form of program code (i.e., instructions) embodied in tangible media, such as floppy diskettes, CD-ROMs, hard drives, or any other machine-readable storage medium, wherein, when the program code is loaded into and executed by a machine, such as a computer, the machine becomes an apparatus for practicing the invention. In the case of program code execution on programmable computers, the

computer will generally include a processor, a storage medium readable by the processor (including volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device. One or more programs are preferably implemented in a high level procedural or object oriented programming language to communicate with a computer system. However, the program(s) can be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language, and combined with hardware implementations.

[0083] The methods and apparatus of the present invention may also be embodied in the form of program code that is transmitted over some transmission medium, such as over electrical wiring or cabling, through fiber optics, or via any other form of transmission, wherein, when the program code is received and loaded into and executed by a machine, such as an EPROM, a gate array, a programmable logic device (PLD), a client computer, a video recorder or the like, the machine becomes an apparatus for practicing the invention. When implemented on a general-purpose processor, the program code combines with the processor to provide a unique apparatus that operates to perform the indexing functionality of the present invention.

[0084] While the present invention has been described in connection with the preferred embodiments of the various figures, it is to be understood that other similar embodiments may be used or modifications and additions may be made to the described embodiment for performing the same function of the present invention without deviating there from. For example, while exemplary embodiments of the invention are described in the context of digital devices emulating the functionality of personal computers and PDAs, one skilled in the art will recognize that the present invention is not limited to such digital devices, as described in the present application may apply to any number of existing or emerging computing devices or environments, such as a gaming console, handheld computer, portable computer, etc. whether wired or wireless, and may be applied to any number of such computing devices connected via a communications network, and interacting across the network. Furthermore, it should be emphasized that a variety of computer platforms, including handheld device operating systems and other application specific operating systems, are herein contemplated, especially as the number of wireless networked devices continues to proliferate. Therefore, the present invention should not be limited to any single embodiment, but rather construed in breadth and scope in accordance with the appended claims.